Measuring Ineffectiveness

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
ellen.voorhees@nist.gov

ABSTRACT

An evaluation methodology that targets ineffective topics is needed to support research on obtaining more consistent retrieval across topics. Using average values of traditional evaluation measures is not an appropriate methodology because it emphasizes effective topics: poorly performing topics' scores are by definition small, and they are therefore difficult to distinguish from the noise inherent in retrieval evaluation. We examine two new measures that emphasize a system's worst topics. While these measures focus on different aspects of retrieval behavior than traditional measures, the measures are less stable than traditional measures and the margin of error associated with the new measures is large relative to the observed differences in scores.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*

General Terms

Measurement, Experimentation

Keywords

evaluation, worst-case behavior

1. INTRODUCTION

An important component of effectiveness for operational retrieval systems is the ability of the system to return reasonable results for every topic. Unfortunately, the standard retrieval evaluation paradigm based on average scores is not an effective tool to help system builders improve the effectiveness of their worst performing topics. Average scores are dominated by the better performing topics, so optimizing MAP or precision at 10 documents retrieved tends to make the effective topics even more effective while not improving the performance of the least effective topics. An evaluation methodology that specifically targets ineffective topics is needed to support research on obtaining more consistent retrieval across topics.

Reliably evaluating the effectiveness of poorly performing topics is difficult, however. Retrieval evaluation is known to be a noisy process that is highly dependent on the particular topics used in the test set. By definition, the evaluation scores for poorly performing topics are small. "Significant" improvement in scores this small is generally indistinguishable from the noise inherent in retrieval evaluation. Reducing the noise by using more topics requires a very large pool of poorly performing topics.

This poster uses the results of the TREC 2004 Robust Retrieval track [2] to explore these issues. Retrieval results submitted to the track were evaluated using both standard measures and two new measures that emphasize poorly performing topics. The results confirm that the new measures focus attention on a system's worst-performing topics. The new measures are less stable than traditional measures, though, and the margin of error associated with the measures is large relative to the observed differences between scores.

2. EVALUATING POOR PERFORMANCE

The task within the Robust Retrieval track was a traditional ad hoc task using a total of 100 topics. One set of 50 topics was drawn from old TREC topics and were selected because most systems performed poorly on the topic in its original TREC ad hoc task. A second set of 50 topics was created for the track. A run submitted to the track consisted of a ranked list of the top 1000 documents retrieved for each of the 100 topics. Seventy-eight runs from 16 different participating groups were submitted to the track.

Two new measures were defined in the track. The first measure was the percentage of topics that retrieved no relevant documents in the top ten retrieved. If one accepts "no relevant documents in the top ten retrieved" as an adequate definition of poorly performing topic, then this is a direct measure of the behavior of interest and is therefore a very intuitive and easily understood measure. However, it is a very coarse measure. The second measure, currently known as the "area measure", is based on MAP scores where the average is computed using only the worst X topics rather than all topics in the test set. Denote the average score over the worst X topics as $\text{MAP}(X)$, and let Q be the number of topics in the test set. Then the value of the area measure is the area under the curve when $\text{MAP}(X)$ is plotted against X for $X = 1 \ldots Q/4$.

3. ANALYSIS OF MEASURES

Table 1 gives evaluation scores for one run for each of the top 10 groups as determined by MAP scores for runs that

This paper is authored by an employee(s) of the United States Government and is in the public domain.
SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

Table 1: Evaluation scores for description-only runs.

Tag	MAP	P10	%no	area
pircRBd2	0.290	0.457	8	0.022
uwmtCR0	0.276	0.453	11	0.018
aplrob03d	0.273	0.393	22	0.008
humR03de	0.263	0.418	19	0.009
VTDokrcgp5	0.256	0.408	15	0.008
fub03InOLe3	0.252	0.413	17	0.008
UIUC03Rd3	0.250	0.390	21	0.006
Sel78QE	0.243	0.372	25	0.003
THUIRr0305	0.243	0.410	12	0.015
SABIR03BF	0.226	0.342	23	0.006

Table 2: Correlation between system rankings.

	P10	% no	area
MAP	0.592	0.180	0.584
P10		0.397	0.493
% no			0.457

used only the description part of the topic statement. The table gives the scores for each of four measures as computed when using the entire set of 100 topics. The four measures are MAP (MAP), the average of precision at 10 documents retrieved (P10), the percentage of topics with no relevant in the top 10 retrieved (%no), and the area measure (area).

The difference in the magnitudes of the MAP and area scores illustrates why in practice optimizing MAP scores does not improve poorly performing topics. Since the magnitude of the MAP score is so much larger than an individual poorly performing topic's average precision score, a poorly performing topic would have to improve dramatically to have any noticeable affect on the MAP score

The runs in Table 1 are ordered by decreasing MAP score. The order would be different if the runs were sorted by a different measure. For example, the THUIRr0305 run would be ranked third rather than ninth if the runs were sorted by the area measure. The similarity between different measures can be quantified using the Kendall τ correlation between the system rankings produced by the respective measures as shown in Table 2. These correlations are quite low, supporting the contention that the measures are affected by different aspects of retrieval behavior.

While the correlation between system rankings produced by different measures does not directly address the issue of the quality of the respective measures, in practice MAP is an effective, stable measure so very low correlations with MAP can be a sign of a flawed measure. To further analyze the behavior of the two new measures, we performed two tests for estimating the stability of an evaluation measure [1, 3]. One test computes an error rate for an evaluation measure by counting how often the measure disagrees with respect to which of two systems being compared is preferred. The second test computes a characteristic difference in evaluation scores required to have 95% confidence in the decision that one run is better than another.

The error rate was computed using 1000 different test sets of 50 topics each that were created by randomly selecting 50 topics from the set of 100 topics used in the track. For all pairs of runs A and B, we counted the number of test sets for which A evaluated as better than B ($A > B$), B evaluated as better than A ($B > A$), and A and B evaluated as equivalent ($A = B$). (Two runs were considered equivalent if the difference in their scores was less than 5 % of the larger score.) The error rate is defined as the sum over all run pairs of the smaller of $A > B$ and $B > A$, divided by the total number of comparisons. The proportion of ties, a measure of the discrimination power of the evaluation measure, is the sum over all pairs of $A = B$ divided by the total number of comparisons.

Table 3: Error rate and proportion of ties for different measures.

	Error Rate (%)	Proportion of Ties
MAP	1.4	0.171
P10	2.6	0.224
% no	9.1	0.090
area	8.4	0.040

Table 3 shows the error rate and proportion of ties computed for the four different measures. The numbers for MAP and P10 are close to the numbers reported in previous studies. The error rates for the area and %-with-no-relevant measures are substantially greater than for the traditional measures, with the %-with-no-relevant measure having the largest error rate. The proportion of ties for the two new measures is substantially smaller than for the traditional measures.

We computed the characteristic difference for topic set sizes up to 50 topics since they can be directly computed using the 100 topics in the track. For topic sets of size 50, a run needs at least 11 fewer topics with no relevant in the top 10 retrieved to have 95% confidence that it is better than a second run. Only 11.0% of the comparisons across all run pairs and 1000 test sets had a difference at least this large, a small percentage that implies the measure is only able to distinguish grossly different systems. For the area measure, the characteristic difference computed for 95% confidence was 0.025, where the largest area score observed across the 1000 test sets was 0.043. Only 4.6% of the comparisons across all run pairs and the 1000 test sets had a difference in area score greater than 0.025, an even smaller percentage that implies the vast majority of systems are indistinguishable from one another.

4. CONCLUSION

Reliably measuring the effectiveness of poorly performing topics is important but inherently difficult. Our attempt to focus on ineffective topics was successful in that it introduced new measures that emphasize systems' worst topics. However, the evaluation is very coarse. The new measures are defined over a subset of topics in the test set, causing them to be much less stable than than traditional measures for a given test set size. In turn, the instability causes the margin of error associated with the measures to be large relative to the differences in scores commonly observed.

5. REFERENCES

[1] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of SIGIR 2000*, pages 33–40, 2000.

[2] Ellen M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *Proceedings of TREC 2003*, 2004. To appear.

[3] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of SIGIR 2002*, pages 316–323, 2002.